Basic Info:

Language : Python

Modules :

A. Pandas
B. Sci kit learn
C. fbprophet
D. Matplotlib

Model/Algorithm : **ARIMA/Prophet (time series)**

Transaction data

Before Parsing:

```
RangeIndex: 545648 entries, 0 to 545647
Data columns (total 9 columns):
customer_id         545648 non-null object
tran_id             545648 non-null int64
tran_date           545648 non-null object
tran_amount         545648 non-null float64
merchant_name       545648 non-null object
merchant_country    545648 non-null object
merchant_city       545648 non-null object
mcc_code            545648 non-null int64
card_id             545648 non-null object
dtypes: float64(1), int64(2), object(6)
```

```
After Parsing:
```

```
RangeIndex: 545648 entries, 0 to 545647
Data columns (total 10 columns):
index               545648 non-null int64
tran_date           545648 non-null datetime64[ns]
customer_id         545648 non-null object
tran_id             545648 non-null int64
tran_amount         545648 non-null float64
merchant_name       545648 non-null object
merchant_country    545648 non-null object
merchant_city       545648 non-null object
mcc_code            545648 non-null int64
card_id             545648 non-null object
dtypes: datetime64[ns](1), float64(1), int64(3), object(5)
```

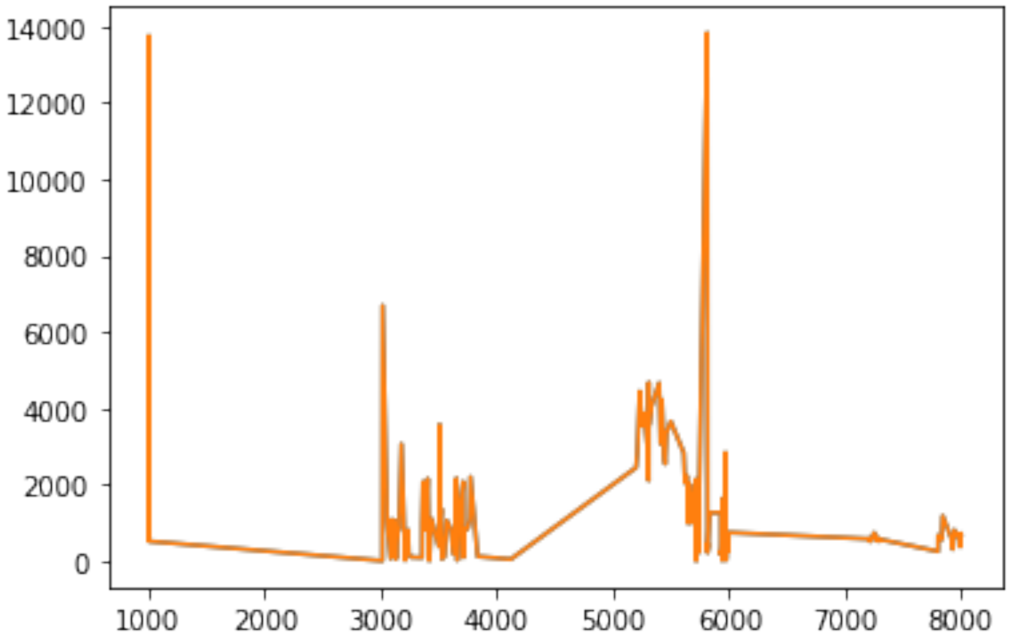
	tran_id	tran_amount	mcc_code
count	5.456480E+05	545648.000000	545648.000000
mean	5.498112E+06	187.387023	5014.344686
std	2.598947E+06	380.893171	1620.966056
min	1.000002E+06	0.000000	1001.000000
25%	3.248337E+06	30.770000	3767.000000
50%	5.494216E+06	55.420000	5499.000000
75%	7.748270E+06	123.460000	5813.000000
max	9.999937E+06	2074.000000	7999.000000

Customer Data

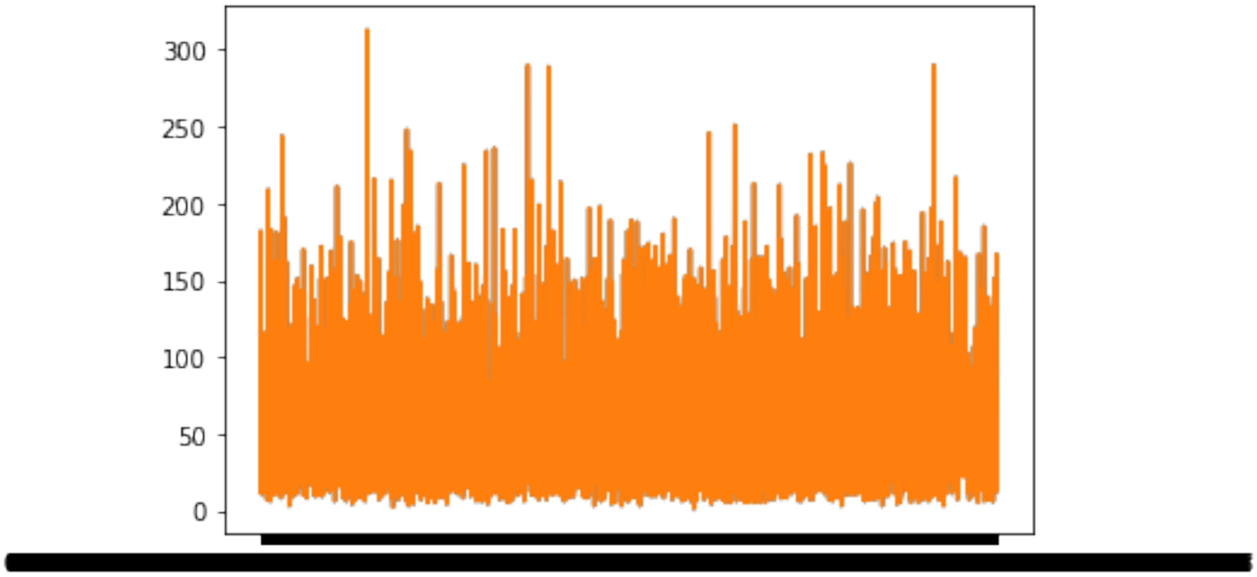
```
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 11 columns):
customer_id         10000 non-null object
card_type           10000 non-null object
card_id             10000 non-null object
mar_status          10000 non-null object
age                 10000 non-null int64
gender              10000 non-null object
customer_country    10000 non-null object
cr_lim_group        10000 non-null int64
customer_city       10000 non-null object
customer_id_new     10000 non-null int64
customer_uid        10000 non-null object
dtypes: int64(3), object(8)
```

	age	cr_lim_group	customer_id_new
count	10000.00000	10000.000000	1.000000E+04
mean	44.391700	14608.953100	1.047848E+07
std	10.071654	12649.626333	2.772524E+05
min	3.000000	5000.000000	1.000002E+07
25%	38.000000	5000.000000	1.023800E+07
50%	45.000000	8155.000000	1.047838E+07
75%	51.000000	21587.500000	1.071296E+07
max	84.000000	82606.000000	1.099999E+07

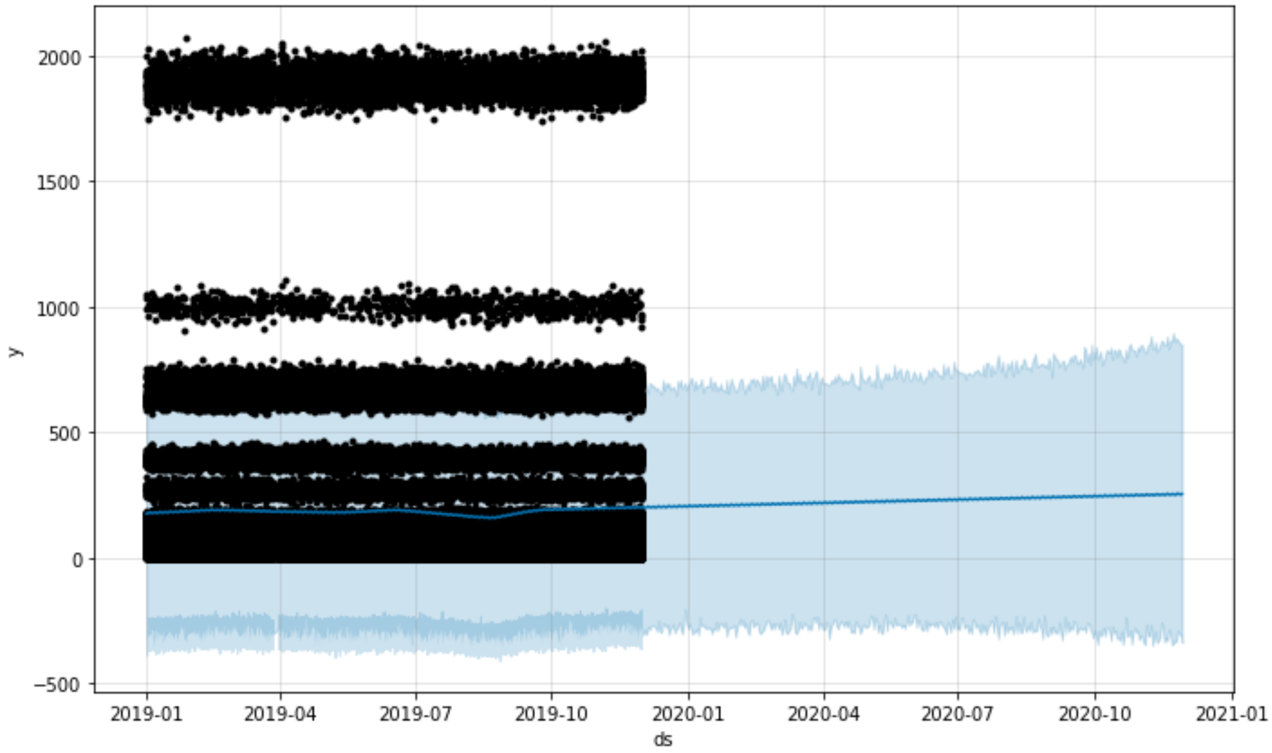
Comparing the number of distinct Customer IDs and Card IDs in both files we can infer that 1000 Customer IDs in customer dataset are not present in transaction dataset.



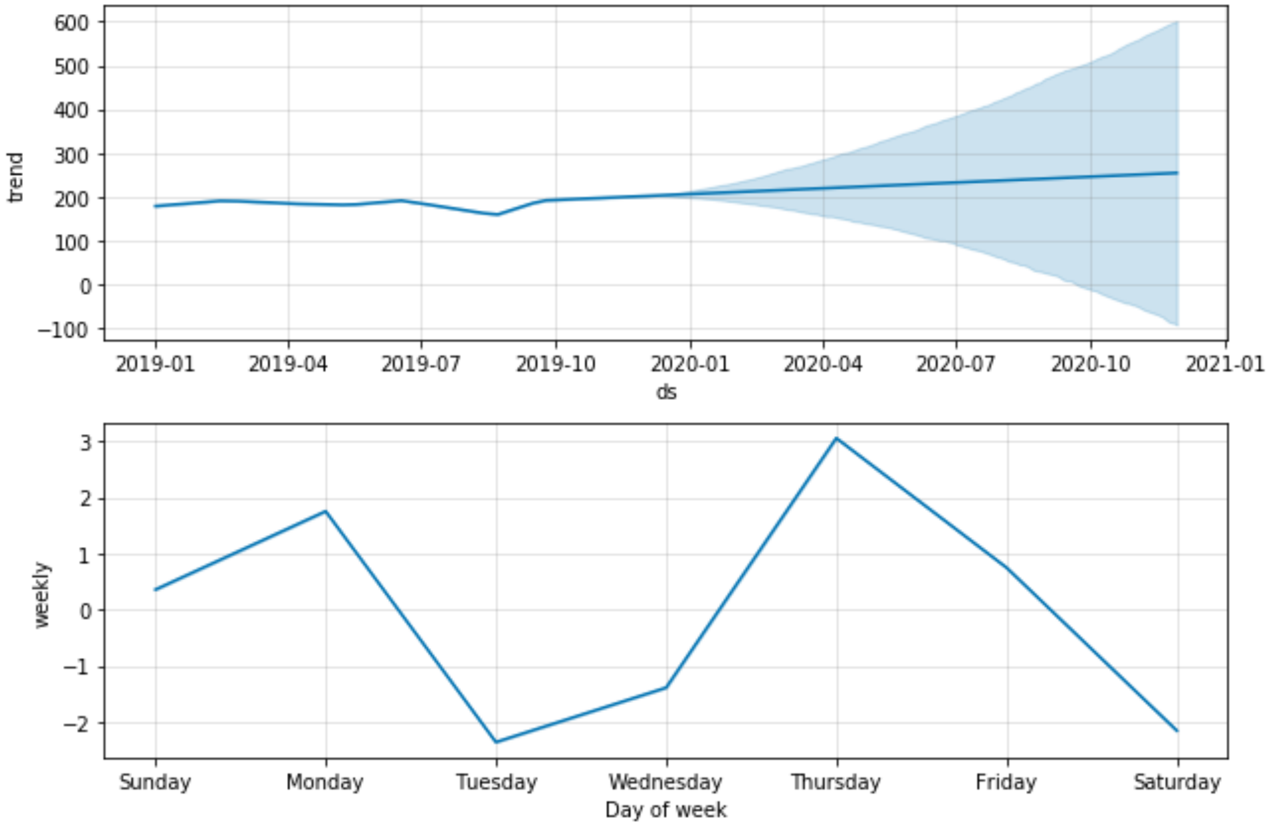
MCC_CODE count



Customer_ID count



Prophet Forecast



Code Implementation:

One of the most critical aspects of any time series algorithm is to ensure stationarity. For any data that is not stationary we need to perform differencing to make it stationary. This is done using specifying the d parameter in ARIMA. We also need to account for seasonality, in a case wherein the data is highly seasonal it is better to use SARIMA.

But before proceeding there are a few steps to be performed such as data preprocessing and any test for checking stationarity like the Dickey-Fuller test.